Indian Institute of Technology Madras, Zanzibar Campus

Z3003 Data Mining: Scaling to Large Data — Mini-Project Evaluation Rubric
BS-DSAI Odd Semester 2026 Due: Friday 20 November 2026, 11:59 pm

Total Marks: 100 Weight: **20% of the final grade** Submission: notebook (.ipynb) + one-page summary (PDF)

- Each section is marked on four levels. Pick the level that best describes the work, then a mark within its band. Marks reward **justified choices**, not the number of techniques used.
- The notebook must run top to bottom. Results that cannot be reproduced from the submitted notebook receive no marks for that section.
- Specific to *your* data beats generic. A sentence that would fit any dataset ("scaling is important") earns little.

---

1 Section 1 - Data Understanding (15 marks)

Criterion	Excellent	Good	Adequate	Weak
Dataset and question (5)	Real, approved dataset; one clear question a decision-maker cares about (5)	Clear question, link to a decision vague (4)	Question stated but broad (2–3)	No clear question (0–1)
Attribute types (5)	Every attribute typed (nominal / binary / ordinal / numeric) with a reason (5)	All typed, few reasons (4)	Most typed (2–3)	Missing or wrong (0–1)
Missing values and quality (5)	Counted per column, cause discussed (random or systematic), quality dimensions checked (5)	Counted and handled; cause not discussed (4)	Counted only (2–3)	Not examined (0–1)

2 Section 2 - Exploratory Data Analysis (20 marks)

Criterion	Excellent	Good	Adequate	Weak
Summaries and plots (6)	Centre, spread and shape reported; plots chosen to answer the question and interpreted (6)	Correct summaries; plots partly interpreted (4–5)	Summaries without interpretation (2–3)	Errors or missing (0–1)
Relationships (7)	Correlation or χ^2 chosen by attribute type, computed correctly, with a named plausible confounder (7)	Correct test; confounder generic (5–6)	Computed, not interpreted (3–4)	Wrong test or missing (0–2)
Distance / similarity (7)	Measure suited to the attribute types, with a worked example on real records (7)	Suitable measure, little interpretation (5–6)	Measure used without justification (3–4)	Missing or unsuitable (0–2)

3 Section 3 - Preprocessing (20 marks)

Criterion	Excellent	Good	Adequate	Weak
Cleaning (6)	Missing values and outliers handled within the right group, effect on the data shown (6)	Sensible method, effect not shown (4–5)	Default method, no reason (2–3)	Not done (0–1)
Transformation (7)	Normalization chosen from the data's shape and **fitted on training data only** (7)	Justified method; fitting order unclear (5–6)	Default method (3–4)	Missing, or leakage (0–2)
Reduction or discretization (7)	PCA (standardized, scree plot, k justified) or entropy discretization; or a reasoned case for neither (7)	Done correctly, weak justification (5–6)	Done, not interpreted (3–4)	Missing or incorrect (0–2)

4 Section 4 - Mining (25 marks)

Criterion	Excellent	Good	Adequate	Weak
Technique choice (5)	One technique (patterns, classification or clustering) that fits the question, with reasons (5)	Fits; reasons thin (4)	Fits loosely (2–3)	Does not fit (0–1)
Correct application (8)	Parameters justified (min_sup, depth, k, $\varepsilon\ldots$); a baseline included (8)	Correct, defaults mostly (6–7)	Runs with errors of method (3–5)	Incorrect (0–2)
Evaluation (8)	Right measure for the task (lift + Kulczynski; precision / recall / F1 / AUC with CV; silhouette + stability), honest about limits (8)	Right measure, no validation (6–7)	Accuracy or confidence only (3–5)	Not evaluated (0–2)
Interpretation (4)	Rules, trees or clusters explained in plain words with their meaning for the question (4)	Partly explained (3)	Listed only (1–2)	None (0)

5 Section 5 - Findings and Discussion (20 marks)

Criterion	Excellent	Good	Adequate	Weak
One-page summary (8)	A non-expert could act on it: question, answer, numbers, recommendation, limits (8)	Clear, little action (6–7)	Technical or vague (3–5)	Missing (0–2)

Scaling (6)	What changes at 1000× the data, with a specific method (sampling, FP-Growth, mini-batch, SON…) (6)	Specific but brief (4–5)	Generic (2–3)	Missing (0–1)
Ethics and limits (6)	A risk specific to this data (privacy, fairness by group, misuse) and a mitigation (6)	Specific risk, no mitigation (4–5)	Generic (2–3)	Missing (0–1)

6 Deductions and Policies

- **Late:** no late submissions are accepted. Work submitted after the deadline receives zero.
- **Data leakage** (test data used to fit scalers, choose k or tune parameters): Section 3 transformation and Section 4 evaluation capped at "Adequate".
- **Not reproducible:** the section(s) whose results cannot be regenerated score 0.
- **Unapproved dataset** (no checkpoint submission on 23 October): −5 marks.
- **Academic integrity:** uncited code, data or AI-generated text is treated under the institute's policy; state AI tool use in a final "Tools used" cell.

Marker's Grading Sheet

Student: _____ Roll No: _____ Dataset: _____

Criterion	Max	Score	Comment
1. Dataset and question	5		
1. Attribute types	5		
1. Missing values and quality	5		
2. Summaries and plots	6		
2. Relationships	7		
2. Distance / similarity	7		
3. Cleaning	6		
3. Transformation (no leakage)	7		
3. Reduction or discretization	7		
4. Technique choice	5		
4. Correct application	8		
4. Evaluation	8		
4. Interpretation	4		
5. One-page summary	8		
5. Scaling	6		
5. Ethics and limits	6		
Subtotal	**100**		
Deductions (late, leakage, reproducibility, approval)			
Final mark	**100**		
Contribution to grade = final × 0.20	**20**		

Strongest aspect: _____

Most important improvement: _____

Marker: _____ Date: _____